Synthetic text extruding machines: A linguist-eye view on their narrow range of applicability

Emily M. Bender
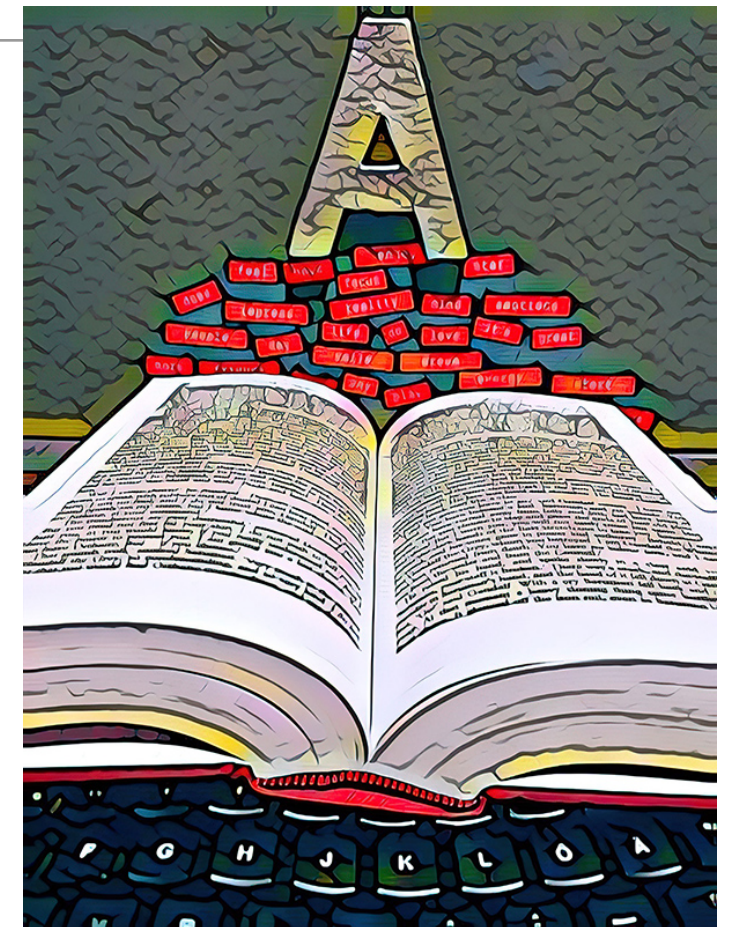University of Washington
@emilymbender
@emilymbender@dair-community.social

UCLA Linguistics
May 17, 2024

The applications of large language models seem endless

New Meta AI demo writes racist and inaccurate scientific literature, gets pulled

Galactica language model generated convincing text about fact and nonsense alike.

02-21-23

A science fiction magazine closed submissions after being bombarded with stories written by ChatGPT

In a case of life (or something) imitating art, an award-winning publisher of science fiction says it's being overrun with AI-generated work.

The applications of large language models seem endless

A news site used AI to write articles. It was a journalistic disaster.

The tech site CNET sent a chill through the media world when it tapped artificial intelligence to produce surprisingly lucid news stories. But now its human staff is writing a lot of corrections.

BOT BUST

Professor Flunks All His Students After ChatGPT Falsely Claims It Wrote Their Papers

Texas A&M University–Commerce seniors who have already graduated were denied their diplomas because of an instructor who incorrectly used AI software to detect cheating

The applications of large language models seem endless

Here's What Happens When Your Lawyer Uses ChatGPT

A lawyer representing a man who sued an airline relied on artificial intelligence to help prepare a court filing. It did not go well.

SHOTS - HEALTH NEWS

National Eating Disorders Association phases out human helpline, pivots to chatbot

May 31, 2023 · 5:08 PM ET

FROM

Artificial intelligence (AI)

US eating disorder helpline takes down AI chatbot over harmful advice

The applications of large language models seem endless

Department of Homeland Security Unveils Artificial Intelligence Roadmap, Announces Pilot Projects to Maximize Benefits of Technology, Advance Homeland Security Mission

Release Date: March 18, 2024

- HSI will test AI to enhance investigative processes focused on detecting fentanyl and increasing efficiency of investigations related to combatting child sexual exploitation.

- FEMA will deploy AI to help communities plan for and develop hazard mitigation plans to build resilience and minimize risks.

- USCIS will use AI to improve immigration officer training.

Outline

- Brief overview & history of language models

- Form vs. meaning: Why language models don't "understand"

- The race for scale: On the dangers of stochastic parrots

- Use cases for synthetic text

- Directions forward (regulation, combatting AI hype)

What's a language model?

- Better term: "corpus model" (Veres 2022)

- Given a collection of text (corpus) representing a language, how likely is a given string to appear?

- Earliest were n-gram language models (Shannon 1948)

  - Unigram: relative frequency of single words

  - Bigram: relative frequency of words given one previous word

  - Trigram: relative frequency of words give two previous words

What are language models good for?

- Ranking spelling correction candidates

- Ranking acoustic model outputs in automatic transcription

- Ranking translation model outputs in machine translation

- Simplified text entry (T9)

What's a neural language model?

- So-called "neural nets" are not artificial brains/minds

- Collections of "perceptrons": Mathematical model based on a simplified version of 1940s understanding of neurons
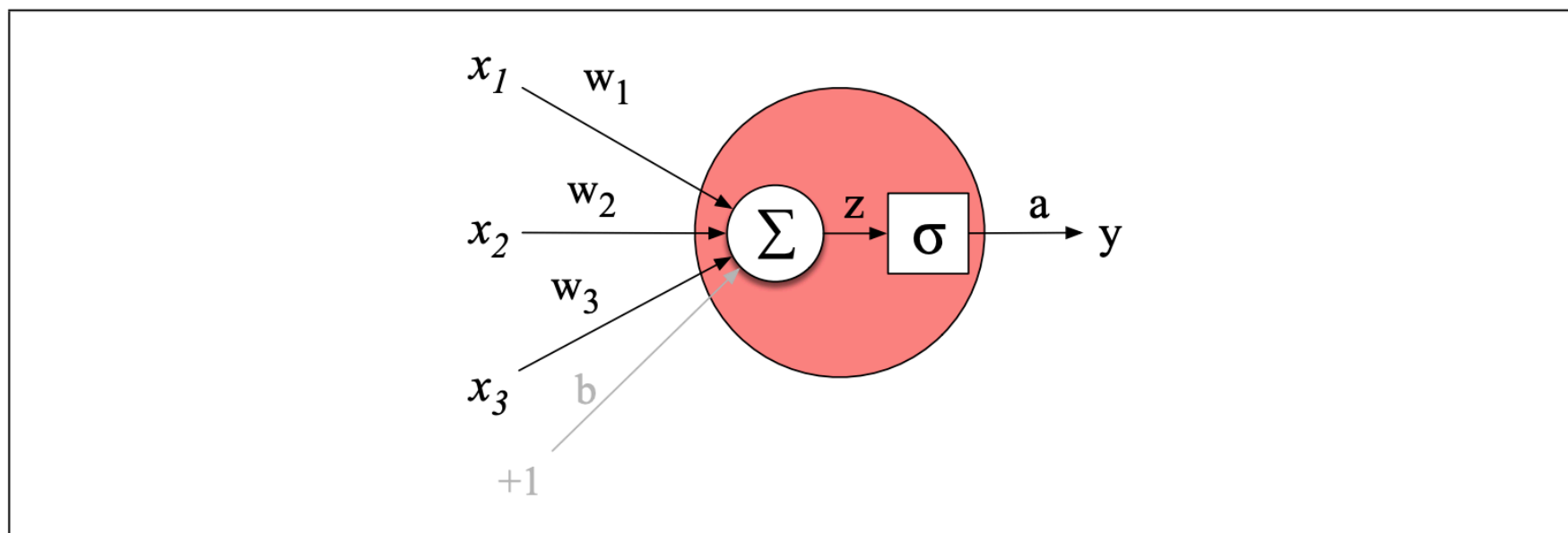
Figure 7.2 A neural unit, taking 3 inputs x_1, x_2, and x_3 (and a bias b that we represent as a weight for an input clamped at +1) and producing an output y. We include some convenient intermediate variables: the output of the summation, z, and the output of the sigmoid, a. In this case the output of the unit y is the same as a, but in deeper networks we'll reserve y to mean the final output of the entire network, leaving a as the activation of an individual node.

(Jurafsky & Martin 2023, Ch 7)

What's a neural language model?

- "Neural net" whose input is a sequence of words and output is a probability distribution over the vocabulary — how likely is each word to come next?

- Trained with "back propagation": compare actual next word to predictions and, when different, adjust weights throughout the network (slightly) (Bengio et al 2003)

- Represent words as "embeddings" (dense vectors reflecting word co-occurrence) rather than character strings, for better generalization across words (Mikolov et al 2013)

- Performance improvement through architecture innovations like Long Short-Term Memory (Hochreiter and Schmidhuber, 1997) and Transformer (Vaswani et al 2017) models and training paradigms (BERT; Devlin et al 2017)
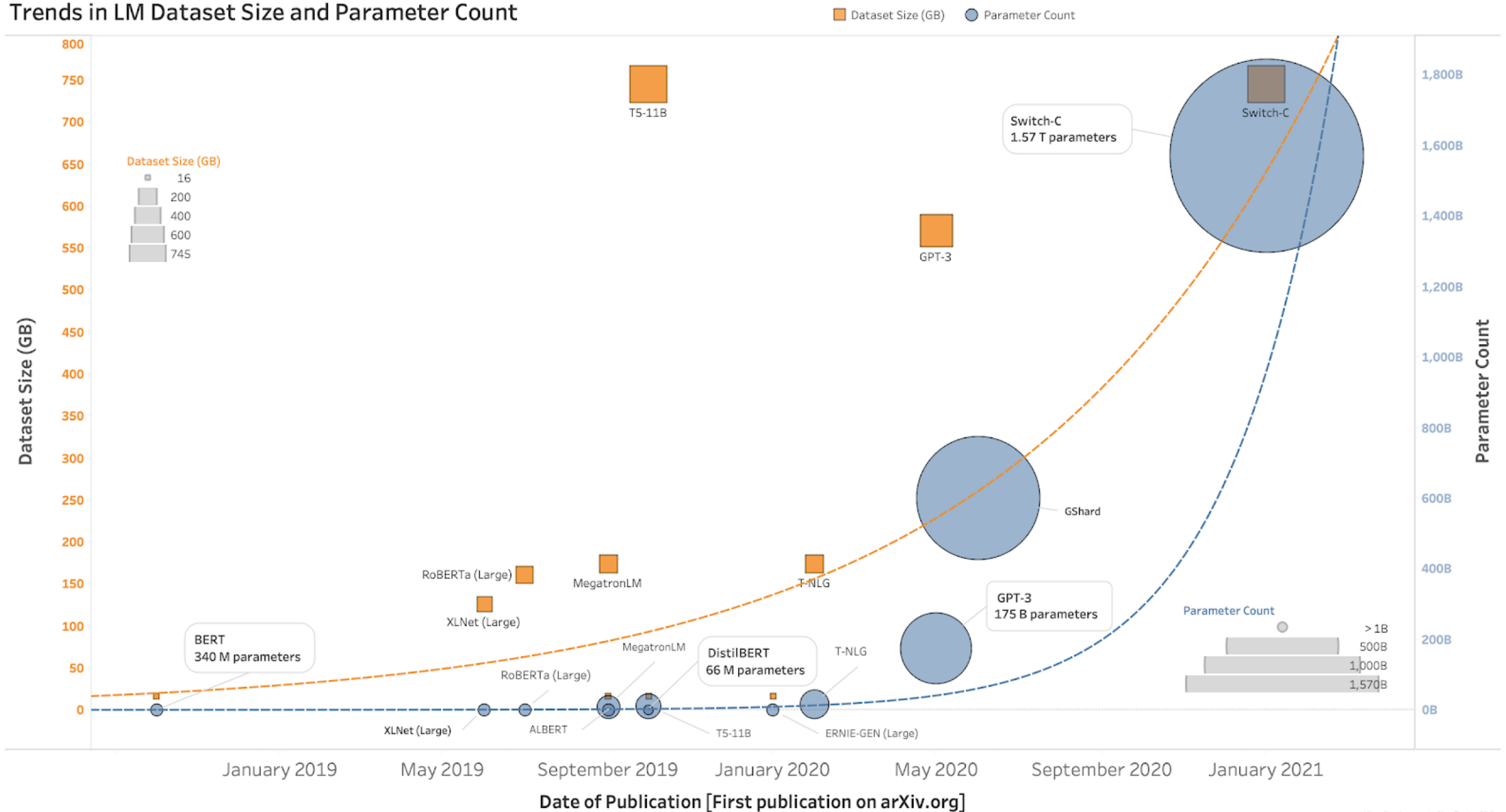
What are neural language models good for?

- Much smoother automatic transcription and machine translation output

- Query expansion in search

- Grammar checker

- Autocorrect

- Word "embeddings" => dramatic improvements to almost every kind of language technology
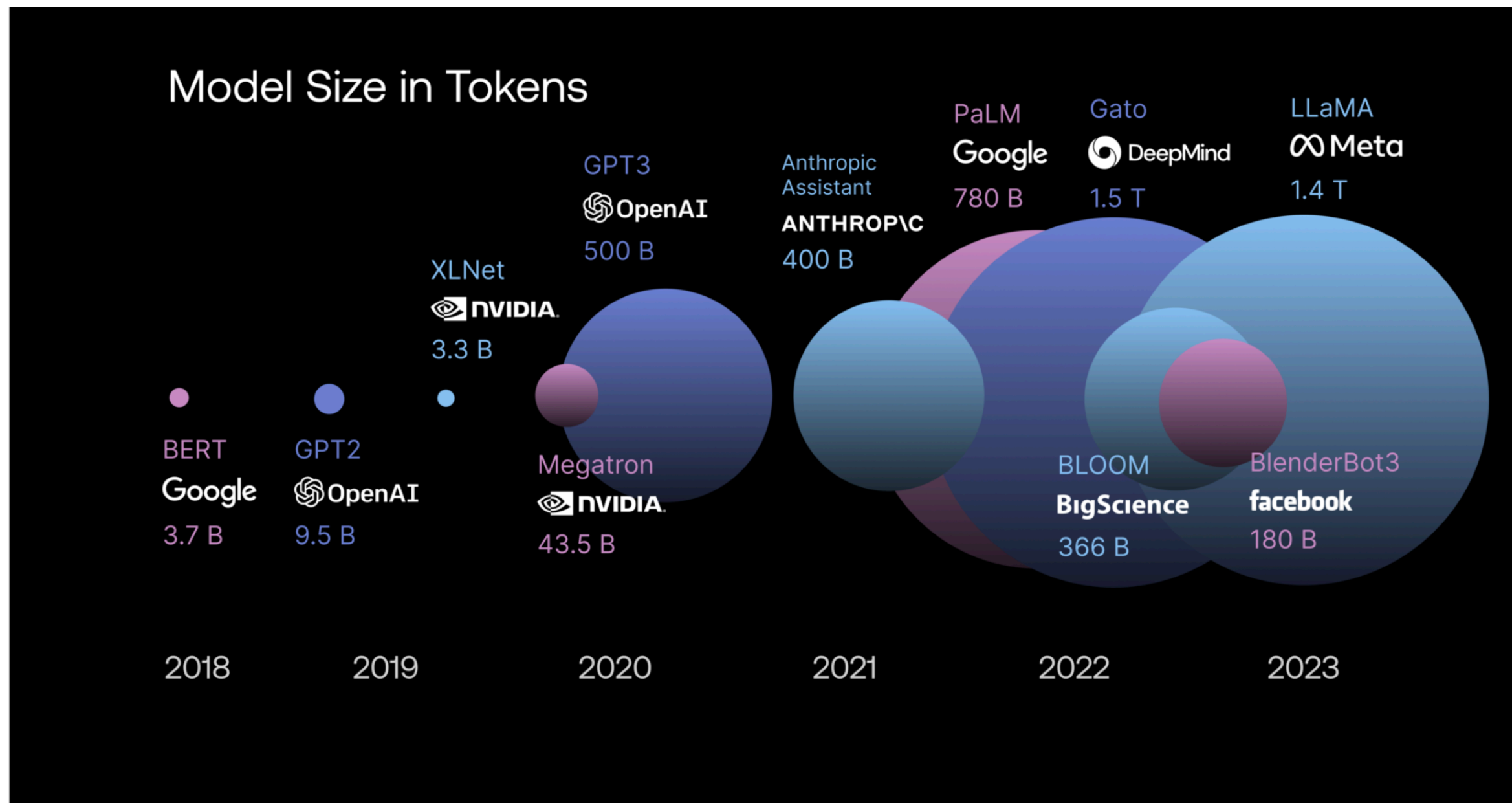
What's a large language model?

(Bender, Gebru et al 2021; design by Denise Mak)

What's a large language model?

Model Size in Tokens

Model	Provider	Size
BERT	Google	3.7 B
GPT2	OpenAI	9.5 B
XLNet	NVIDIA	3.3 B
Megatron	NVIDIA	43.5 B
GPT3	OpenAI	500 B
Anthropic Assistant	ANTHROP\C	400 B
PaLM	Google	780 B
Gato	DeepMind	1.5 T
LLaMA	Meta	1.4 T
BLOOM	BigScience	366 B
BlenderBot3	facebook	180 B

2018   2019   2020   2021   2022   2023

https://scale.com/guides/large-language-models

How about 2024?
How large is Gemini 1.5's training data?

Data overview	
Training Dataset	Gemini 1.5 Pro is trained on a variety of multimodal and multilingual data. Refer to Section 3 (Training Infrastructure and Dataset).
Evaluation Dataset	Refer to Section 4 (Evaluation) where we cover core capability evaluations in text, audio and vision.
Fine-tuning Dataset	Gemini 1.5 Pro is fine-tuned on a collection of multimodal data containing paired instructions and corresponding desired responses. Refer to Section 3 (Training Infrastructure and Dataset) for more.

3. Training Infrastructure and Dataset

Like Gemini 1.0 Ultra and 1.0 Pro, Gemini 1.5 Pro is trained on multiple 4096-chip pods of Google's TPUv4 accelerators, distributed across multiple datacenters, and on a variety of multimodal and multilingual data. Our pre-training dataset includes data sourced across many different domains, including web documents and code, and incorporates image, audio, and video content. For the instruction-tuning phase we finetuned Gemini 1.5 Pro on a collection of multimodal data (containing paired instructions and appropriate responses), with further tuning based on human preference data. We refer readers to the Gemini 1.0 Technical Report (Gemini-Team et al., 2023) for further information.

(Gemini Team, 2024)

What are large language models good for?

- Automatic transcription, machine translation

- "End-to-end" approaches to many, many language technology tasks:

  - Summarization

  - Sentiment analysis

  - Taking multiple-choice tests

  - …

What is "generative AI"?

- Turning systems meant for classification/ranking inside-out

- Instead of "Which string is more plausible?" we get "What word comes next?"

- Cover term for other kinds of synthetic media machines (audio, image, video) as well

- Not "AI", and definitely not "AGI"

What is "generative AI" good for?

When, if ever, is synthetic text safe, appropriate, and desirable?

Outline

- Brief overview & history of language models

- Form vs. meaning: Why language models don't "understand"

- The race for scale: On the dangers of stochastic parrots

- Use cases for synthetic text

- Directions forward (regulation, combatting AI hype)

Climbing towards NLU:
On Meaning, Form, and Understanding
in the Age of Data

Emily M. Bender, University of Washington
Alexander Koller, Saarland University

ACL 2020

BERT fanclub

- "In order to train a model that **understands** sentence relationships, we pre-train for a binarized next sentence prediction task that can be trivially generated from any monolingual corpus." (Devlin et al 2019)

- "Using BERT, a pretraining language model, has been successful for single-turn machine **comprehension** …" (Ohsugi et al 2019)

- "The surprisingly strong ability of these models to **recall factual knowledge** without any fine-tuning demonstrates their potential as unsupervised open-domain QA systems." (Petroni et al 2019)

What is meaning?

- Competent speakers easily conflate 'form' and 'meaning' because we can only rarely perceive one without the other

- In order to understand what's going on with ChatGPT we need to take a closer look

Working definitions

- **Form** : marks on a page, pixels or bytes, movements of the articulators

- **Meaning** : relationship between linguistic form and something external to language

  - $M \subseteq E \times I$: pairs of expressions and communicative intents

  - $C \subseteq E \times S$: pairs of expressions and their standing meanings

- **Understanding** : given an expression *e*, in a context, recover the communicative intent *i*

BERTology

- Strand 1: What are BERT and similar learning about language structure?

  - Distributional similarities between words (Lin et al 2015, Mikolov et al 2013)

  - Something analogous to dependency structure (Tenney et al 2019, Hewitt & Manning 2019)

- Strand 2: What information are the Transformers using to 'beat' the tasks?

  - Niven & Kao (2019): in ARCT, BERT is exploiting spurious artifacts

  - McCoy et al (2019): in NLI, BERT leans on lexical, subsequence, & constituent overlap heuristics

- Our contribution: Theoretical perspective on why models exposed only to form can never learn meaning

So how do babies learn language?

- Interaction is key: Exposure to a language via TV or radio alone is not sufficient (Snow et al 1976, Kuhl 2007)

- Interaction allows for joint attention: where child and caregiver are attending to the same thing and mutually aware of this fact (Baldwin 1995)

- Experimental evidence shows that more successful joint attention leads to faster vocabulary acquisition (Tomasello & Farrar 1986, Baldwin 1995, Brooks & Meltzoff 2005)

- Meaning isn't in form; rather, languages are rich, dense ways of providing cues to communicative intent (Reddy 1979). Once we learn the systems, we can use them in the absence of co-situatedness.

Thought experiment: Meaning from form alone
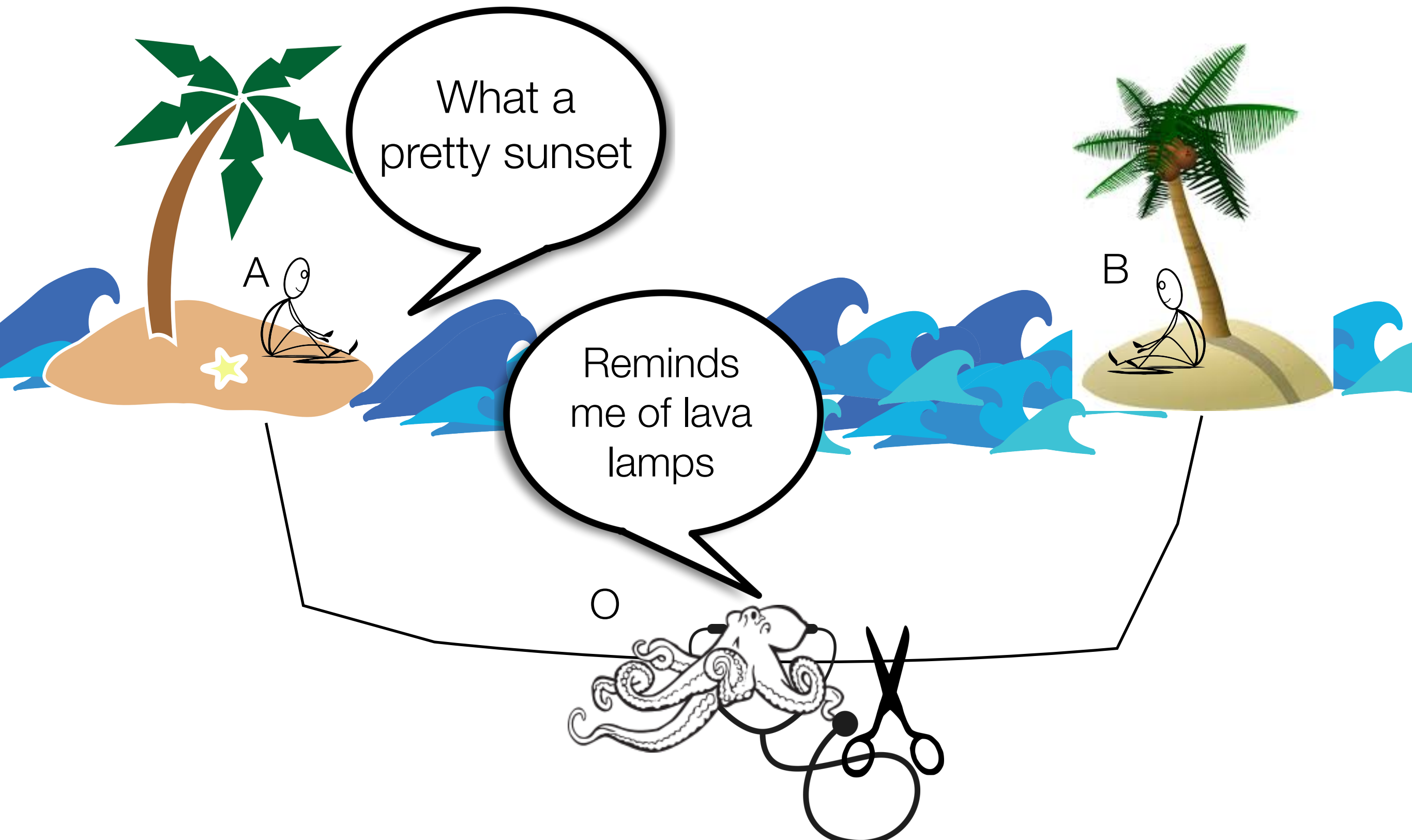
Thought experiment: Meaning from form alone
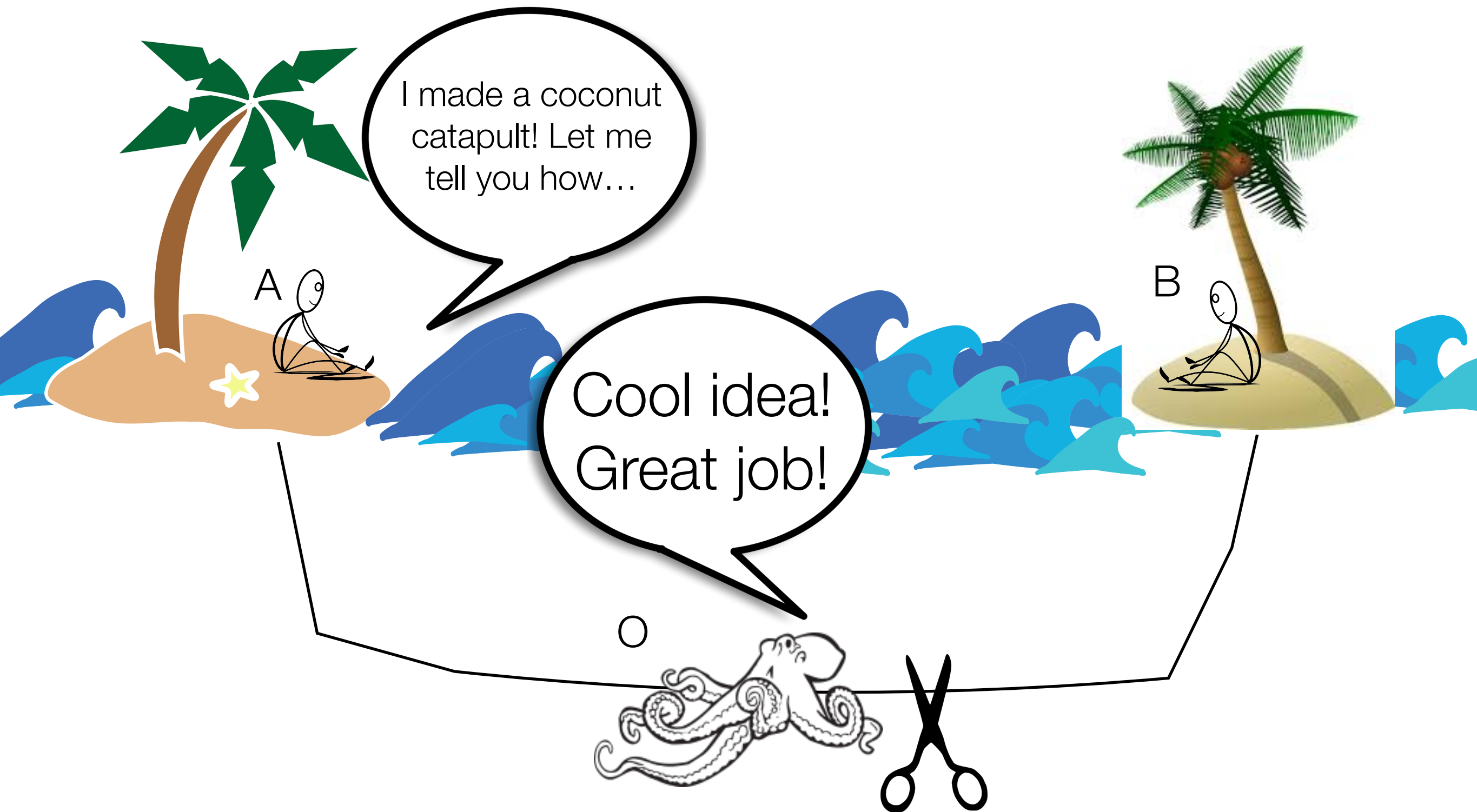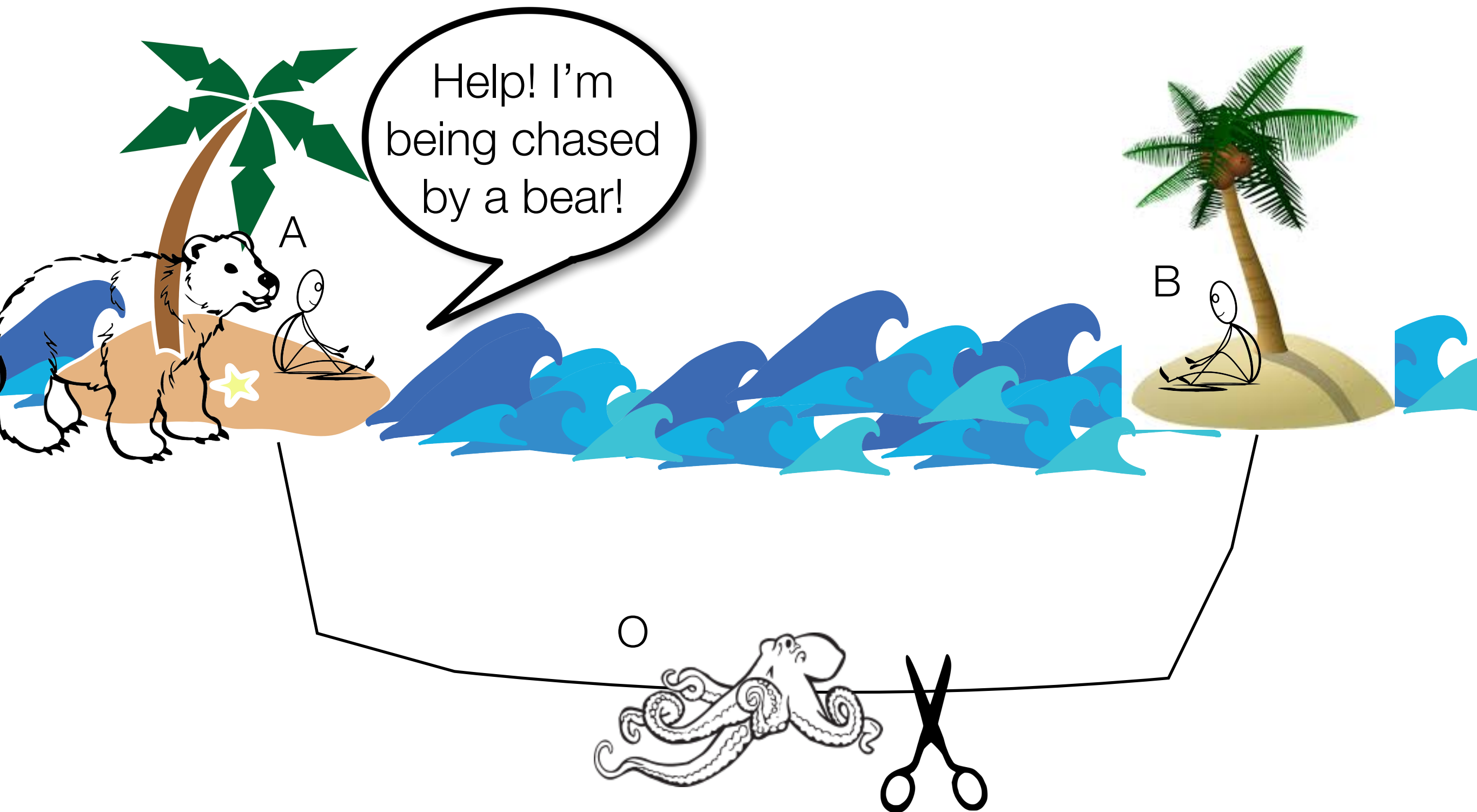
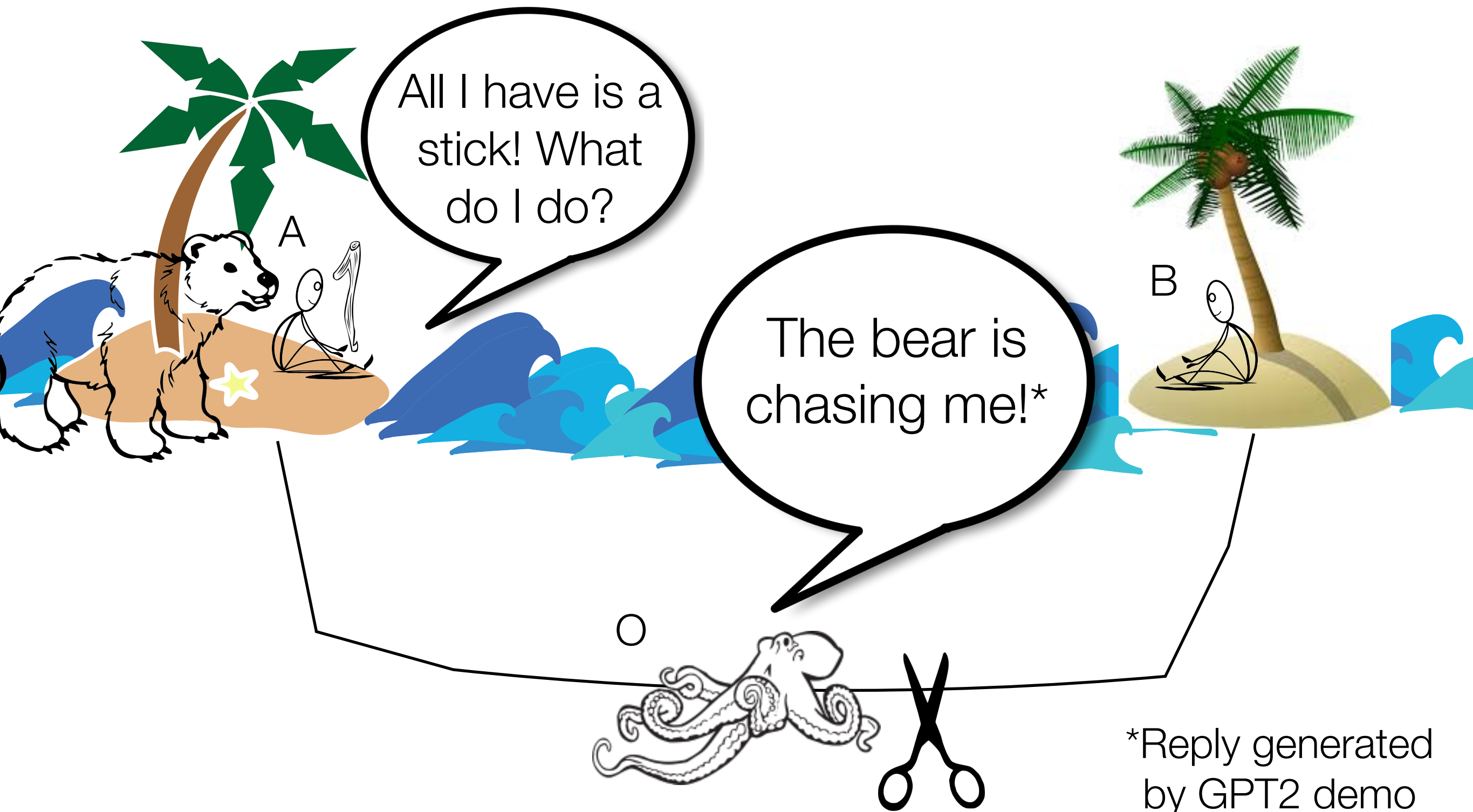# Thought experiment: Meaning from form alone

Thought experiment: Meaning from form alone

All I have is a stick! What do I do?

The bear is chasing me!*

A

B

O

*Reply generated by GPT2 demo

Thought experiment: Meaning from form alone
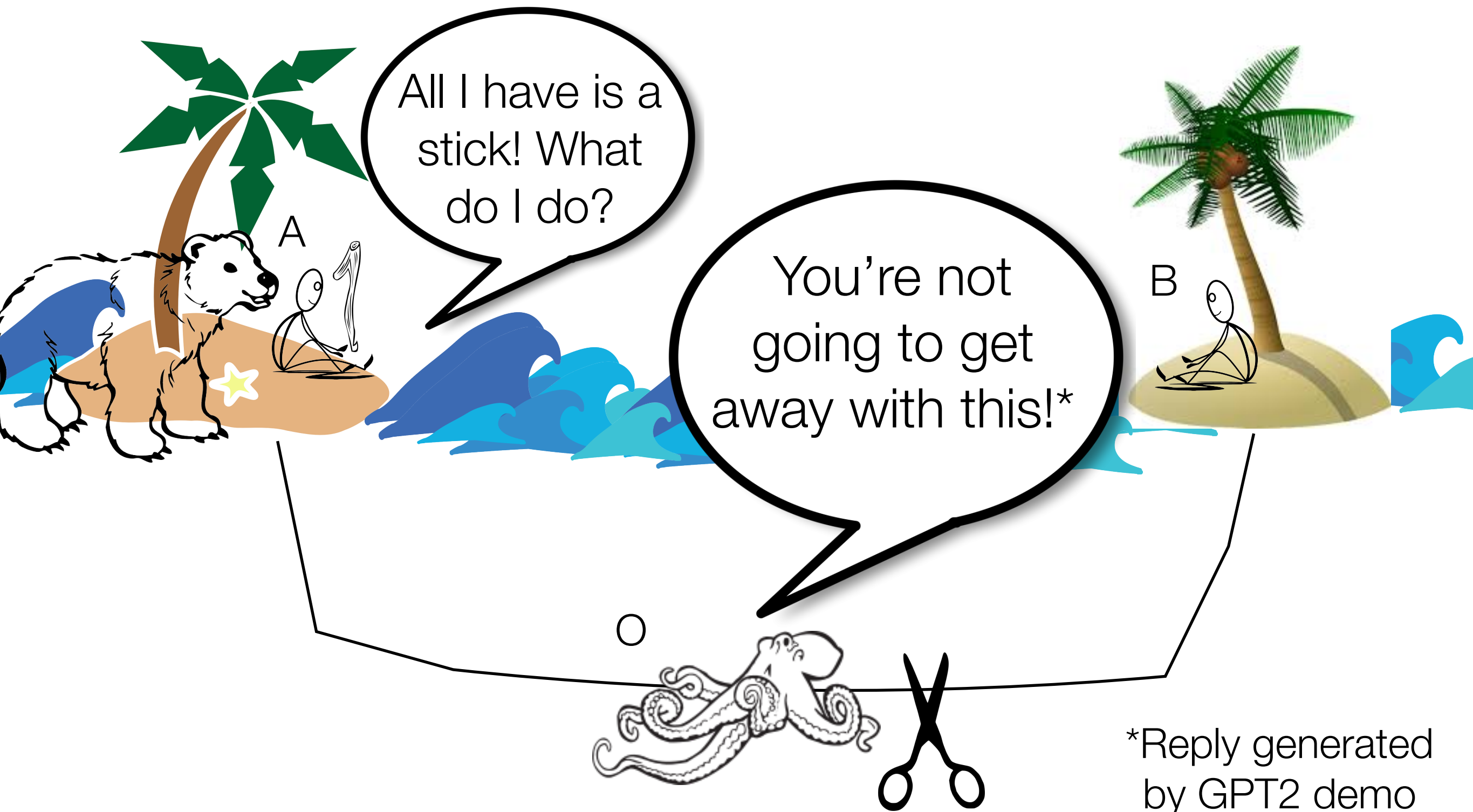
Octopus Test: Analysis

- O did not learn to communicate successfully, and the reason is that O did not learn meaning.

- This is because O could only observe forms, and meaning can't be learned from form alone.

  Learning the meaning relation requires access to the outside world so communicative intents can be hypothesized and tested.

- To the extent that A finds O's utterances meaningful,
  it was not because O's utterances made sense;
  it is because A, as a human active listener, *could make sense of them*.

2023 update: National Library of Thailand

- You're in the National Library of Thailand

- Unlimited time, unlimited delicious Thai food, no people to interact with

- All documents with images or non-Thai text removed

- Can you learn Thai?

- How?

(Photo credit:
Pat Roengpitya)

2023 update: National Library of Thailand

- Look for illustrated encyclopedia or scientific articles with English words (sorry, these were removed)

- Find common subsequences, deduce that these are function morphemes

- Look for a book that is obviously a translation of a book you know well

- Relax & eat yummy Thai food

- => Only strategies that bring in external information work
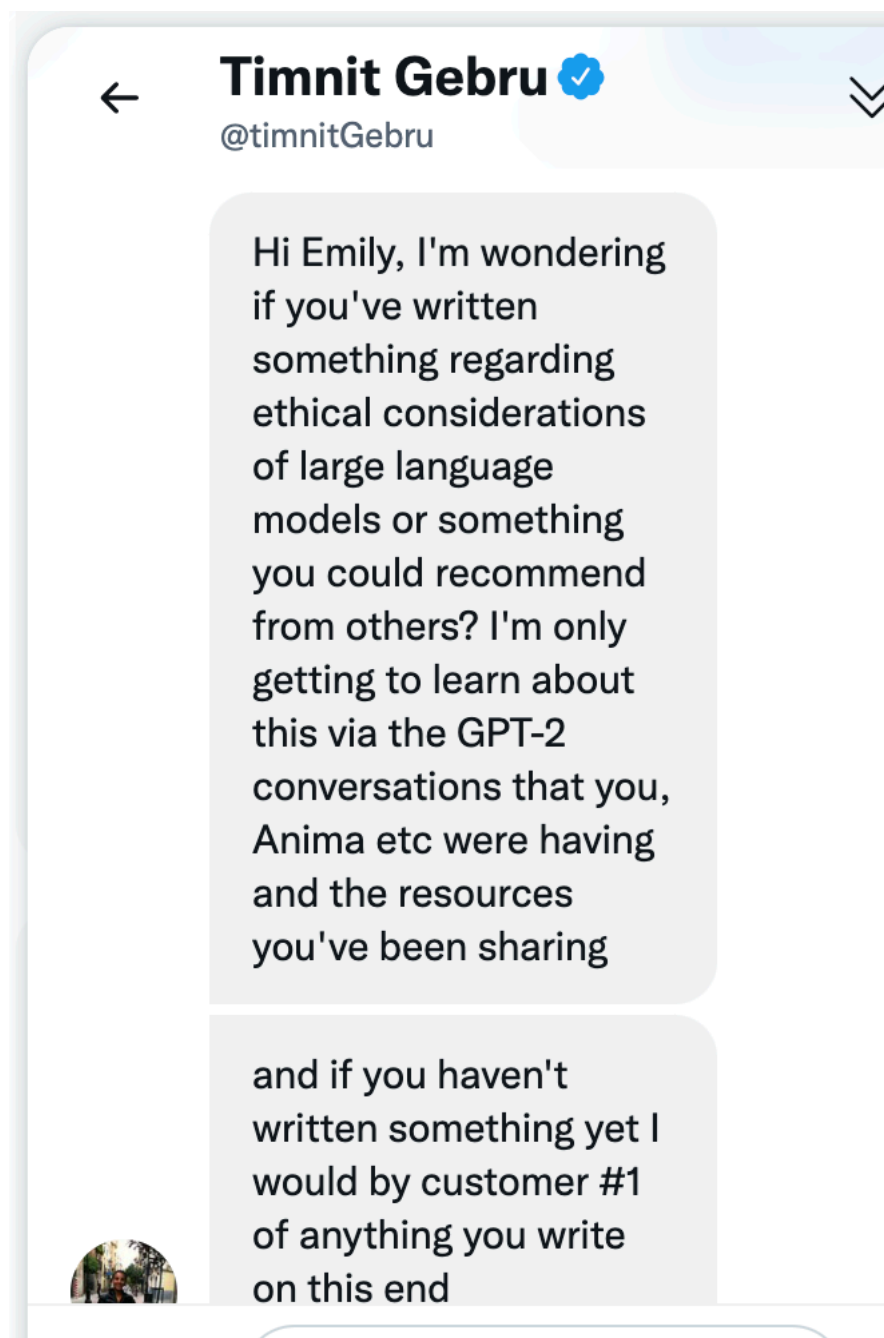

(Photo credit:
Pat Roengpitya)

Outline

- Brief overview & history of language models

- Form vs. meaning: Why language models don't "understand"

- The race for scale: On the dangers of stochastic parrots

- Use cases for synthetic text

- Directions forward (regulation, combatting AI hype)

2020: Origin story of the Stochastic Parrots paper

- Started off as a Twitter DM conversation, with Dr. Timnit Gebru:

Timnit Gebru ✓
@timnitGebru

Hi Emily, I'm wondering if you've written something regarding ethical considerations of large language models or something you could recommend from others? I'm only getting to learn about this via the GPT-2 conversations that you, Anima etc were having and the resources you've been sharing

and if you haven't written something yet I would by customer #1 of anything you write on this end

This is somethign I'm trying to advocate for at Google

sent them some of your tweets as well haha

so many converstations about how we're not leading in large language models and should

GPT-3 is so impressive etc

and each time I'm like ANNDD see what Emily has to say

I can think of three other angles around ethical implications of GPT-3 and the like:

1) Carbon cost of creating the damn things (see Strubell et al at ACL 2019)
2) AI hype/people claiming it's understanding when it isn't (Bender & Koller at ACL 2020)
3) Deepfakes/random generated text that no one is accountable for but which is interpreted as meaningful.

Sep 8, 2020, 4:54 PM ✓

Bender, Gebru et al 2021
On the Dangers of Stochastic Parrots:
Can Language Models be too big? 🦜

- *Prabhakaran*: Prabhakaran et al 2012, Prabhakaran & Rambow 2017, Hutchison et al 2020

- *Hutchinson*: Hutchinson et al 2019, 2020, 2021

- *Díaz*: Lazar et al 2017, Díaz et al 2018

We would like you to consider

- Are ever larger language models (LMs) inevitable or necessary?

- What costs are associated with this research direction and what should we consider before pursuing it?

- Do the field of natural language processing or the public that it serves in fact need larger LMs?

- If so, how can we pursue this research direction while mitigating its associated risks?

- If not, what do we need instead?

What are the risks?

Environmental costs & financial inaccessibility

Environmental and financial costs

- Average human across the globe responsible for 5t of CO_2 emissions per year*

- Strubell et al. (2019)

  - Transformer model training procedure on GPUs 284t of CO_2 emissions

  - 0.1 BLUE score increase en-de results in increase of ~$150,000 in compute cost

  - Encourage reporting training time and sensitivity to hyperparameters

  - Suggest more equitable access to compute clouds through government investment

- Which researchers and which languages get to 'play' in this space and who is cut out?

*Source: Our World In Data

Current mitigation efforts

- Renewable energy sources

  - Still incur a cost on the environment & take away from other potential uses of green energy

- Prioritize computationally efficient hardware

  - SustainNLP workshop

  - Green AI and promoting efficiency as evaluation metric (Schwartz et al 2020)

- Document energy and carbon metrics

  - Energy Usage Reports (Lottick et al 2019)

  - Experiment-impact-tracker (Henderson et al 2020)

Costs and risks to whom?

- Large LMs, particularly those in English and other high-resource languages, benefit those who have the most in society

- Marginalized communities around the world impacted most by climate change

  - Maldives threatened by rising sea levels (Anthoff et al 2010)

  - 800,000 residents of Sudan affected by flooding (7/2020-10/2020)*

- But these communities are rarely able to see benefits of language technology because LLMs aren't built for their languages, Dhivehi and Sudanese Arabic

*Source: https://www.aljazeera.com/news/2020/9/25/over-800000-affected-in-sudan-flooding-un

What are the risks?

Unmanageable training data

A large dataset is not necessarily diverse

- Who has access to the Internet and is contributing?

  - Younger people and those from developed countries

- Who is being subject to moderation?

  - Twitter - accounts receiving death threats more likely to be suspended than those issuing threats (see also Marshall 2021)

- What parts of the Internet are being scraped?

- Reddit - US users 67% men and 64% are ages 18-29 (Pew)

- Wikipedia - only 8.8-15% are women or girls

- Not sites with fewer incoming and outgoing links, like blogs

- Who is being filtered out?

  - Filtering lists primarily target words referencing sex, likely also filtering LGBTQ online spaces (see also Dodge et al 2021)

Static data/Changing social views

- LMs run the risk of 'value lock', reifying older, less-inclusive understandings

- BLM movement lead to increased number of articles on shootings of Black people and past events were also documented and updated (Twyman et al 2017)

  - But media also doesn't cover all events and tend to focus on more dramatic content

- LMs encode hegemonic views; retraining/fine-tuning would require thoughtful curation (see Solaiman and Dennison 2021 for partial proof of concept)

- See also Birhane 2021: ML applied as prediction is inherently conservative

Bias

- Research in probing LMs for bias has provided a wealth of examples of bias

  - See Blodgett et al 2020 for a critical overview

- Documentation of the problem is an important first step, but not a solution

- Automated processing steps may themselves be unreliable

- Probing requires knowing what social categories the LM may be biased against

  - Need for local input before deployment

Curation, documentation, accountability

- *How big is too big?*

  - Budget for documentation and only collect as much data as can be documented

  - Documentation: understand sources of bias & potential mitigating strategies

  - No documentation: potential for harm without recourse

- *Documentation debt*: datasets both undocumented and too big to document post-hoc

What are the risks?

Research trajectories

Research time is a valuable resource

- Focus on LMs and achieving new SOTA on leaderboards, particularly NLU

- But LMs have been shown to excel due to spurious dataset artifacts (Niven & Kao 2019, Bras et al 2020)

- LMs trained only on linguistic form don't have access to meaning (Bender & Koller 2020)

- Are we actually learning about machine language understanding? Building effective, useable, well-scoped technology?

What are the risks?

Potential harms of synthetic language

We can't help ourselves

- Human-human interaction is co-constructed and leads to a shared model of the world (Reddy 1979, Clark 1996)

- Text generated by an LM is not grounded in any communicative intent, model of the world, or model of the reader's state of mind

- Counter-intuitive, given the increasing fluency of text synthesis machines, but:

  - Have to account for our predisposition to interpret locutionary artifacts as conveying coherent meaning & intent (Weizenbaum 1976, Nass et al 1994)

Stochastic 🦜

- An LM is a system for haphazardly stitching together linguistic forms from its vast training data, without any reference to meaning: a *stochastic parrot*.

- Nonetheless, humans encountering synthetic text make sense of it

  - Coherence is in the eye of the beholder

Potential harms

- Denigration, stereotype threat, hate speech:
  harms to reader, harms to bystanders

- Cheap synthetic text can boost extremist recruiting (McGuffie & Newhouse 2020)

- LM errors attributed to human author in MT (better fluency reads as better accuracy)

- LMs as hidden components can influence query expansion & results (Noble 2018)

Potential harms

- These harms largely stem from the interaction of the ersatz fluency of today's language models + human tendency to attribute meaning to text

- Deeply connected to issue of accountability:

  - Synthetic text can enter conversations without anyone being accountable for it

- Accountability key to responsibility for truthfulness and to situating meaning

- Maggie Nelson (2015): "Words change depending on who speaks them; there is no cure."

Stochastic Parrots coda (2024 update)

- "How do you feel now that your predictions have come true?"

- Those weren't predictions, they were warnings!

- What we didn't predict/notice at the time:

  - Exploitative labor practices

  - Just how enthusiastic people would be about synthetic text

  - Pollution of the information ecosystem

  - The transition to treating LLMs as "everything machines", i.e. an "unscoped technology" (Gebru & Torres 2023)

Outline

- Brief overview & history of language models

- Form vs. meaning: Why language models don't "understand"

- The race for scale: On the dangers of stochastic parrots

- Use cases for synthetic text

- Directions forward (regulation, combatting AI hype)

What is "generative AI" good for?

When, if ever, is synthetic text safe, appropriate, and desirable?

Is information access ("search") a good use case?

- No: Large language models are designed to make stuff up

- Arguably more dangerous at 95% accurate than 70%

- More importantly:

  - Exacerbates authoritativeness and thus the problems identified by Noble (2018)

  - Cuts off user from sense making processes in information access

(Shah & Bender 2022, 2024)

Criteria for a good use case

- What matters is language form (content is unimportant)

  - OR: Content can efficiently and effectively be thoroughly vetted

- Ersatz fluency and coherence would not be misleading

- Problematic biases and hateful content can be identified and filtered

- Originality is not required (risk of plagiarism is minimized)

- … and you are using an LLM created with fair labor practices and without data theft

Candidate use cases potentially meeting those criteria

• Dialogue partner in language learning scenario

• Non-player characters in interactive games

• Short-form writing support (email, press releases)

• "What's that called?" type searches

=> Are these worth the environmental & social costs?

Safe use of text synthesis machines

- Access to clear and thorough documentation of training data

  - Bender & Friedman 2018, Bender et al 2021, Gebru et al 2021, Mitchell et al 2019, Hinds et al 2018, Chmielinski et al 2022

- Software is thoroughly tested for intended use case

  - And is known to be of a stable version that won't change behind the scenes

- Use of text synthesis is clearly indicated

  - Especially any text published without thorough vetting

- Accountability for content (and originality) clearly held by a person or organization of people

Outline

- Brief overview & history of language models

- Form vs. meaning: Why language models don't "understand"

- The race for scale: On the dangers of stochastic parrots

- Use cases for synthetic text

- Directions forward (regulation, combatting AI hype)

Steps forward: Individual

- Be a critical consumer of news about "AI" systems and the systems themselves

  - What's the task the tech is meant to be used for?

  - How was it evaluated in that context?

  - What data was used for system development?

  - Who is benefitting from the tech?

  - What actions are being attributed to the tech that might be better understood as actions by people (using automation)?

Steps forward: Combatting AI Hype

- https://www.dair-institute.org/maiht3k/

Steps forward: Societal

- Apply existing regulation vigorously: as the <u>FTC has said</u>, there is no "AI loophole"

- Insist on transparency:

  - Of training data (what's in it, where was it taken from)

  - Of presence of synthetic media (watermarking)

- Insist that accountability rest with people not machines

- Insist on recourse

- Insist on labor rights

Outline

Thank you!

- Brief overview & history of language models

- Form vs. meaning: Why language models don't "understand"

- The race for scale: On the dangers of stochastic parrots

- Use cases for synthetic text

- Directions forward (regulation, combatting AI hype)

References

Anthoff, D., Nicholls, R. J., and Tol, R. S. (2010). The economic impact of substantial sea-level rise. *Mitigation and Adaptation Strategies for Global Change*, 15(4):321–335.

Baldwin, D. A. (1995). Understanding the link between joint attention and language. In Moore, C. and Dunham, P. J., editors, *Joint Attention: Its Origins and Role in Development*, pages 131–158. Psychology Press.

Bender, E. M., Freidman, B., and McMillan-Major, A. (2021a). A guide for writing data statements for natural language processing.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S., and et al (2021b). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT 2021*.

Bender, E. M. and Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):11371155.

Birhane, A. (2021). The Impossibility of Automating Ambiguity. *Artificial Life*, 27(1):44–61.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Bras, R. L., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M. E., Sabharwal, A., and Choi, Y. (2020). Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*.

Brooks, R. and Meltzoff, A. N. (2005). The development of gaze following and its relation to language. *Developmental Science*, 8(6):535–543.

Clark, H. H. (1996). *Using Language*. Cambridge University Press, Cambridge.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Díaz, M., Johnson, I., Lazar, A., Piper, A. M., and Gergle, D. (2018). Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 114. Association for Computing Machinery, New York, NY, USA.

Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., and Gardner, M. (2021). Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for datasets. *Commun. ACM*, 64(12):8692.

Gebru, T. and Torres, É. P. (2023). Eugenics and the promise of utopia through artificial general intelligence. Talk presented by Gebru at SaTML 2023. Recording available at `https://www.youtube.com/watch?v=P7XT4TWLzJw`.

Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Hind, M., Mehta, S., Mojsilovic, A., Nair, R. G., Ramamurthy, K. N., Olteanu, A., and Varshney, K. R. (2018). Increasing trust in ai services through supplier's declarations of conformity. *IBM J. Res. Dev.*, 63:6:1–6:13.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9:1735–1780.

Hutchinson, B., Pittl, K. J., and Mitchell, M. (2019). Interpreting social respect: A normative lens for ML models. *CoRR*, abs/1908.07336.

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., and Denuyl, S. (2020). Social biases in NLP models as barriers for persons with disabilities. *CoRR*, abs/2005.00813.

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., and Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 560575, New York, NY, USA. Association for Computing Machinery.

Jurafsky, D. and Martin, J. H. (2023). *Speech and Language Processing (3rd ed. draft)*. Available from `https://web.stanford.edu/~jurafsky/slp3/`.

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science*, 10(1):110–120.

Lazar, A., Díaz, M., Brewer, R., Kim, C., and Piper, A. M. (2017). Going gray, failure to hire, and the ick factor: Analyzing how older bloggers talk about ageism. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 655668, New York, NY, USA. Association for Computing Machinery.

Lin, C.-C., Ammar, W., Dyer, C., and Levin, L. (2015). Unsupervised POS induction with word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1311–1316, Denver, Colorado. Association for Computational Linguistics.

Lottick, K., Susai, S., Friedler, S. A., and Wilson, J. P. (2019). Energy usage reports: Environmental awareness as part of algorithmic accountability. In *Proceedings of Workshop on Tackling Climate Change with Machine Learning, NeurIPS 2019*, Vancouver, Canada.

Marshall, B. (2021). Algorithmic misogynoir in content moderation practice. `https://us.boell.org/en/2021/06/21/algorithmic-misogynoir-content-moderation-practice-1`.

McCoy, T., Pavlick, E., and Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

McGuffie, K. and Newhouse, A. (2020). The radicalization risks of GPT-3 and advanced neural language models. Technical report, Center on Terrorism, Extremism, and Counterterrorism, Middlebury Institute of International Studies at Monterrey. `https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf`.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. ACM.

Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78.

Nelson, M. (2015). *The Argonauts*. Graywolf Press, Minneapolis.

Niven, T. and Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Ohsugi, Y., Saito, I., Nishida, K., Asano, H., and Tomita, J. (2019). A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 11–17, Florence, Italy. Association for

Computational Linguistics.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Prabhakaran, V. and Rambow, O. (2017). Dialog structure through the lens of gender, gender environment, and power. *Dialogue & Discourse*, 8(2):21–55.

Prabhakaran, V., Rambow, O., and Diab, M. (2012). Predicting overt display of power in written dialogs. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 518–522, Montréal, Canada. Association for Computational Linguistics.

Reddy, M. J. (1979). The conduit metaphor: A case of frame conflict in our language about language. In *Metaphor and Thought*, pages 164–201.

Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., baptiste Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A., Millican, K., Dyer, E., Glaese, M., Sottiaux, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Molloy, J., Chen, J., Isard, M., Barham, P., Hennigan, T., McIlroy, R., Johnson, M., Schalkwyk, J., Collins, E., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Meyer, C., Thornton, G., Yang, Z., Michalewski, H., Abbas, Z., Schucher, N., Anand, A., Ives, R., Keeling, J., Lenc, K., Haykal, S., Shakeri, S., Shyam, P., Chowdhery, A., Ring, R., Spencer, S., Sezener, E., Vilnis, L., Chang, O., Morioka, N., Tucker, G., Zheng, C., Woodman, O., Attaluri, N., Kocisky, T., Eltyshev, E., Chen, X., Chung, T., Selo, V., Brahma, S., Georgiev, P., Slone, A., Zhu, Z., Lottes, J., Qiao, S., Caine, B., Riedel, S., Tomala, A., Chadwick, M., Love, J., Choy, P., Mittal, S., Houlsby, N., Tang, Y., Lamm, M., Bai, L., Zhang, Q., He, L., Cheng, Y., Humphreys, P., Li, Y., Brin, S., Cassirer, A., Miao, Y., Zilka, L., Tobin, T., Xu, K., Proleev, L., Sohn, D., Magni, A., Hendricks, L. A., Gao, I., Ontan, S., Bunyan, O., Byrd, N., Sharma, A., Zhang, B., Pinto, M., Sinha, R., Mehta, H., Jia, D., Caelles, S., Webson, A., Morris, A., Roelofs, B., Ding, Y., Strudel, R., Xiong, X., Ritter, M., Dehghani, M., Chaabouni, R., Karmarkar, A., Lai, G., Mentzer, F., Xu, B., Li, Y., Zhang, Y., Paine, T. L., Goldin, A., Neyshabur, B., Baumli, K., Levskaya, A., Laskin, M., Jia, W., Rae, J. W., Xiao, K., He, A., Giordano, S., Yagati, L., Lespiau, J.-B., Natsev, P., Ganapathy, S., Liu, F., Martins, D., Chen, N., Xu, Y., Barnes, M., May, R., Vezer, A., Oh, J., Franko, K., Bridgers, S., Zhao, R., Wu, B., Mustafa, B., Sechrist, S., Parisotto, E., Pillai, T. S., Larkin, C., Gu, C., Sorokin, C., Krikun, M., Guseynov, A., Landon, J., Datta, R., Pritzel, A., Thacker, P., Yang, F., Hui, K., Hauth, A., Yeh, C.-K., Barker, D., Mao-Jones, J., Austin, S., Sheahan, H., Schuh, P., Svensson, J., Jain, R., Ramasesh, V., Briukhov, A., Chung, D.-W., von Glehn, T., Butterfield, C., Jhakra, P., Wiethoff, M., Frye, J., Grimstad, J., Changpinyo, B., Lan, C. L., Bortsova, A., Wu, Y., Voigtlaender, P., Sainath, T., Smith, C., Hawkins, W., Cao, K., Besley, J., Srinivasan, S., Omernick, M., Gaffney, C., Surita, G., Burnell, R., Damoc, B., Ahn, J., Brock, A., Pajarskas, M., Petrushkina, A., Noury, S., Blanco, L., Swersky, K., Ahuja, A., Avrahami, T., Misra, V., de Liedekerke, R., Iinuma, M., Polozov, A., York, S., van den Driessche, G., Michel, P., Chiu, J., Blevins, R., Gleicher, Z., Recasens, A., Rrustemi, A., Gribovskaya, E., Roy, A., Gworek, W., Arnold, S., Lee, L., Lee-Thorp, J., Maggioni, M., Piqueras, E., Badola, K., Vikram, S., Gonzalez, L., Baddepudi, A., Senter, E., Devlin, J., Qin, J., Azzam, M., Trebacz, M., Polacek, M., Krishnakumar, K., yiin Chang, S., Tung, M., Penchev, I., Joshi, R., Olszewska, K., Muir, C., Wirth, M., Hartman, A. J., Newlan, J., Kashem, S., Bolina, V., Dabir, E., van Amersfoort, J., Ahmed, Z., Cobon-Kerr, J., Kamath, A., Hrafnkelsson, A. M., Hou, L., Mackinnon, I., Frechette, A., Noland, E., Si, X., Taropa, E., Li, D., Crone, P., Gulati, A., Cevey, S., Adler, J., Ma, A., Silver, D., Tokumine, S., Powell, R., Lee, S., Chang, M., Hassan, S., Mincu, D., Yang, A., Levine, N., Brennan, J., Wang, M., Hodkinson, S., Zhao, J., Lipschultz, J., Pope, A., Chang, M. B., Li, C., Shafey, L. E., Paganini, M., Douglas, S., Bohnet, B., Pardo, F., Odoom, S., Rosca, M., dos Santos, C. N., Soparkar, K., Guez, A., Hudson, T., Hansen, S., Asawaroengchai, C., Addanki, R., Yu, T., Stokowiec, W., Khan, M., Gilmer, J., Lee, J., Bostock, C. G., Rong, K., Caton, J., Pejman, P., Pavetic, F., Brown, G., Sharma, V., Lui, M., Samuel, R., Djolonga, J., Mandhane, A., Sjsund, L. L.,

Buchatskaya, E., White, E., Clay, N., Jiang, J., Lim, H., Hemsley, R., Labanowski, J., Cao, N. D., Steiner, D., Hashemi, S. H., Austin, J., Gergely, A., Blyth, T., Stanton, J., Shivakumar, K., Siddhant, A., Andreassen, A., Araya, C., Sethi, N., Shivanna, R., Hand, S., Bapna, A., Khodaei, A., Miech, A., Tanzer, G., Swing, A., Thakoor, S., Pan, Z., Nado, Z., Winkler, S., Yu, D., Saleh, M., Maggiore, L., Barr, I., Giang, M., Kagohara, T., Danihelka, I., Marathe, A., Feinberg, V., Elhawaty, M., Ghelani, N., Horgan, D., Miller, H., Walker, L., Tanburn, R., Tariq, M., Shrivastava, D., Xia, F., Chiu, C.-C., Ashwood, Z., Baatarsukh, K., Samangooei, S., Alcober, F., Stjerngren, A., Komarek, P., Tsihlas, K., Boral, A., Comanescu, R., Chen, J., Liu, R., Bloxwich, D., Chen, C., Sun, Y., Feng, F., Mauger, M., Dotiwalla, X., Hellendoorn, V., Sharman, M., Zheng, I., Haridasan, K., Barth-Maron, G., Swanson, C., Rogoziska, D., Andreev, A., Rubenstein, P. K., Sang, R., Hurt, D., Elsayed, G., Wang, R., Lacey, D., Ili, A., Zhao, Y., Aroyo, L., Iwuanyanwu, C., Nikolaev, V., Lakshminarayanan, B., Jazayeri, S., Kaufman, R. L., Varadarajan, M., Tekur, C., Fritz, D., Khalman, M., Reitter, D., Dasgupta, K., Sarcar, S., Ornduff, T., Snaider, J., Huot, F., Jia, J., Kemp, R., Trdin, N., Vijayakumar, A., Kim, L., Angermueller, C., Lao, L., Liu, T., Zhang, H., Engel, D., Greene, S., White, A., Austin, J., Taylor, L., Ashraf, S., Liu, D., Georgaki, M., Cai, I., Kulizhskaya, Y., Goenka, S., Saeta, B., Vodrahalli, K., Frank, C., de Cesare, D., Robenek, B., Richardson, H., Alnahlawi, M., Yew, C., Ponnapalli, P., Tagliasacchi, M., Korchemniy, A., Kim, Y., Li, D., Rosgen, B., Ashwood, Z., Levin, K., Wiesner, J., Banzal, P., Srinivasan, P., Yu, H., alar nl, Reid, D., Tung, Z., Finchelstein, D., Kumar, R., Elisseeff, A., Huang, J., Zhang, M., Zhu, R., Aguilar, R., Gimnez, M., Xia, J., Dousse, O., Gierke, W., Yeganeh, S. H., Yates, D., Jalan, K., Li, L., Latorre-Chimoto, E., Nguyen, D. D., Durden, K., Kallakuri, P., Liu, Y., Johnson, M., Tsai, T., Talbert, A., Liu, J., Neitz, A., Elkind, C., Selvi, M., Jasarevic, M., Soares, L. B., Cui, A., Wang, P., Wang, A. W., Ye, X., Kallarackal, K., Loher, L., Lam, H., Broder, J., Holtmann-Rice, D., Martin, N., Ramadhana, B., Toyama, D., Shukla, M., Basu, S., Mohan, A., Fernando, N., Fiedel, N., Paterson, K., Li, H., Garg, A., Park, J., Choi, D., Wu, D., Singh, S., Zhang, Z., Globerson, A., Yu, L., Carpenter, J., de Chaumont Quitry, F., Radebaugh, C., Lin, C.-C., Tudor, A., Shroff, P., Garmon, D., Du, D., Vats, N., Lu, H., Iqbal, S., Yakubovich, A., Tripuraneni, N., Manyika, J., Qureshi, H., Hua, N., Ngani, C., Raad, M. A., Forbes, H., Bulanova, A., Stanway, J., Sundararajan, M., Ungureanu, V., Bishop, C., Li, Y., Venkatraman, B., Li, B., Thornton, C., Scellato, S., Gupta, N., Wang, Y., Tenney, I., Wu, X., Shenoy, A., Carvajal, G., Wright, D. G., Bariach, B., Xiao, Z., Hawkins, P., Dalmia, S., Farabet, C., Valenzuela, P., Yuan, Q., Welty, C., Agarwal, A., Chen, M., Kim, W., Hulse, B., Dukkipati, N., Paszke, A., Bolt, A., Davoodi, E., Choo, K., Beattie, J., Prendki, J., Vashisht, H., Santamaria-Fernandez, R., Cobo, L. C., Wilkiewicz, J., Madras, D., Elqursh, A., Uy, G., Ramirez, K., Harvey, M., Liechty, T., Zen, H., Seibert, J., Hu, C. H., Elhawaty, M., Khorlin, A., Le, M., Aharoni, A., Li, M., Wang, L., Kumar, S., Lince, A., Casagrande, N., Hoover, J., Badawy, D. E., Soergel, D., Vnukov, D., Miecnikowski, M., Simsa, J., Koop, A., Kumar, P., Sellam, T., Vlasic, D., Daruki, S., Shabat, N., Zhang, J., Su, G., Zhang, J., Liu, J., Sun, Y., Palmer, E., Ghaffarkhah, A., Xiong, X., Cotruta, V., Fink, M., Dixon, L., Sreevatsa, A., Goedeckemeyer, A., Dimitriev, A., Jafari, M., Crocker, R., FitzGerald, N., Kumar, A., Ghemawat, S., Philips, I., Liu, F., Liang, Y., Sterneck, R., Repina, A., Wu, M., Knight, L., Georgiev, M., Lee, H., Askham, H., Chakladar, A., Louis, A., Crous, C., Cate, H., Petrova, D., Quinn, M., Owusu-Afriyie, D., Singhal, A., Wei, N., Kim, S., Vincent, D., Nasr, M., Choquette-Choo, C. A., Tojo, R., Lu, S., de Las Casas, D., Cheng, Y., Bolukbasi, T., Lee, K., Fatehi, S., Ananthanarayanan, R., Patel, M., Kaed, C., Li, J., Sygnowski, J., Belle, S. R., Chen, Z., Konzelmann, J., Pder, S., Garg, R., Koverkathu, V., Brown, A., Dyer, C., Liu, R., Nova, A., Xu, J., Petrov, S., Hassabis, D., Kavukcuoglu, K., Dean, J., and Vinyals, O. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12):5463.

Shah, C. and Bender, E. M. (2022). Situating search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '22, pages 221–232, New York, NY, USA. Association for Computing Machinery.

Shah, C. and Bender, E. M. (2024). Envisioning information access systems: What makes for good tools

and a healthy web? *ACM Trans. Web.* Just Accepted.

Shannon, C. E. (1948). A mathematical theory of information. *Bell System Technical Journal*, 27:379–423, 623–656.

Snow, C. E., Arlman-Rupp, A., Hassing, Y., Jobse, J., Joosten, J., and Vorster, J. (1976). Mothers' speech in three social classes. *Journal of Psycholinguistic Research*, 5(1):1–20.

Solaiman, I. and Dennison, C. (2021). Process for adapting language models to society (PALMS) with values-targeted datasets. In *NeurIPS 2021*, Sydney, Australia.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Durme, B. V., Bowman, S., Das, D., and Pavlick, E. (2019). What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Tomasello, M. and Farrar, M. J. (1986). Joint attention and early language. *Child Development*, 57(6):1454–1463.

Twyman, M., Keegan, B. C., and Shaw, A. (2017). Black lives matter in wikipedia: Collective memory and collaboration around online social movements. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1400–1412.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*.

Veres, C. (2022). Large language models are not models of natural language: They are corpus models. *IEEE Access*, 10:61970–61979.

Weizenbaum, J. (1976). *Computer Power and Human Reason: From Judgment to Calculation.* Freeman, San Francisco CA.

Sources for news headlines:

- https://arstechnica.com/information-technology/2022/11/after-controversy-meta-pulls-demo-of-ai-model-that-writes-scientific-papers/
- https://www.fastcompany.com/90853591/chatgpt-science-fiction-short-stories-clarkesworld-magazine-submissions
- https://www.washingtonpost.com/media/2023/01/17/cnet-ai-articles-journalism-corrections/
- https://www.rollingstone.com/culture/culture-features/texas-am-chatgpt-ai-professor-flunks-students-false-claims-1234736601/
- https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html
- https://www.npr.org/sections/health-shots/2023/05/31/1179244569/national-eating-disorders-association-phases-out-human-helpline-pivots-to-chatbo
- https://www.theguardian.com/technology/2023/may/31/eating-disorder-hotline-union-ai-chatbot-harm
- https://www.dhs.gov/news/2024/03/18/department-homeland-security-unveils-artificial-intelligence-roadmap-announces

Sources for parrot photos:

- https://www.maxpixel.net/Bird-Red-Parrot-Animal-Fly-Vintage-Wings-1300223
- https://www.maxpixel.net/Parrots-Parrot-Birds-Isolated-Plumage-Branch-Bird-2850879
- https://www.maxpixel.net/Tropical-Animal-World-Bill-Parrot-Cute-Bird-Ara-3080543
- https://www.maxpixel.net/Animal-Ara-Plumage-Isolated-Bird-Parrot-4720084
- https://www.maxpixel.net/Tropical-Ara-Bird-Feather-Exotic-Bill-Parrot-3064137
- https://www.maxpixel.net/Plumage-Colorful-Exotic-Birds-Ara-Parrot-5202301
- https://www.maxpixel.net/Flight-Parrots-Parrot-Isolated-2683451